Содержание

1 Введение **2**

2 Теоретическая часть задания **2**

2.1 Составляющие модели 3

2.2 Тест Дики-Фуллера 4

2.3 Модели типа ARIMA и методология Бокса-Дженкинса . . . 5

2.4 AIC - информационный критерий Акаике 6

3 Используемые библиотеки **7**

4 Реализация программы **7**

Задание №3

Михаил Романцов и Павел Елисеев

1 Введение

Задание №3 заключается в проведении анализа некоторого временного ряда и предсказании его значений для последующих месяцев.

2 Теоретическая часть задания

Широкий круг социально-экономических, технических и физических процессов часто представляется в виде набора последовательностых значений некоторого показателя $Y_1, Y_2, ..., Y_n$, зафиксированных в равноудалённые друг от друга моменты времени. Подобный набор значений $Y_t, t = 1, 2, ..., n$ именуется временным рядом, который представляет собой дискретный временный процесс. В зависимости от свойств различают стационарные и нестационарные временные ряды различных порядков. Так, стационарность второго порядка (слабая стационарность или стационарность в широком смысле) наблюдается, если моменты первого и второго порядка (математическое ожидание, дисперсия и автоковариация) инвариантны по отношению к сдвигу временного аргумента.

Тренд — тенденция изменения показателей временного ряда. Тренды могут быть описаны различными функциями — линейными, степенными, экспоненциальными и т. д.

Сезонность - периодические колебания, наблюдаемые на временных рядах.

Остаток - разница между предсказанным и наблюдаемым значением.

Аддитивная модель имеет вид: Y = T + S + E;

Мультипликативная модель имеет вид: Y = T*S*E;

T - тренд, S - сезональность, E - остаток

При обработке информации о поведении финансовых временных рядов необходимо учитывать, что методы анализа нестационарных случайных процессов существенно отличаются от приёмов работы со стационарынми случайными временными рядами. Однако в рамках системы фондового рынка существуют множество приёмов, описывающих его динамику, которые обладают так называемой однородной нестационарностью и могут быть описаны про помощи подходов, применимых к стационарным рядам. К числу таких методов относится применение линейной стохастической модели авторегрессии и проинтегрированного скользящего среднего ($AutoRegressiveIntegratedMovingAverage, ARIMA$).

В настоящей работе рассматриваются финансовые временные ряды, имеющие стационарность второго порядка. Именно эта степень стационарности, определяемая условиями инвариантности взаимного распределения вероятностей наблюдений, является жёстким условием для построения модели $ARIMA$.

2.1 Составляющие модели

Авторегрессионные модели

Авторегресионная модель порядка p имеет вид

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_p$$

где Y_t - уровень временного ряда в момент времени t (зависимая переменная);

$Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ – уровни временного ряда в моменты времени $t-1, t-2, ..., t-p$, соответственно (независимые переменные);

$\phi_0, \phi_1, \phi_2, ..., \phi_p$ – оцениваемые коэффициенты;

ϵ_t – случайное возмущение, описывающее влияние переменных, не учтенных в модели. Коэффициент ϕ_0 определяет постоянный уровень ряда и связан с математическим ожиданием соотношением $\phi_0 = (1 - \phi_1 - \phi_2 - ... - \phi_p)$.

Модель со скользящим средним

Модель со скользящим средним порядка q задается уравнением

$$Y_t = \epsilon_t - \omega_1 \epsilon_{t-1} - \omega_2 \epsilon_{t-2} - ... - \omega_q \epsilon_{t-q}, \text{ где } Y_t$$

– уровень ряда в момент времени t;

ϵ_{t-i} – значения остатков i временных периодов назад (независимые переменные);

$\omega_1, \omega_2, ..., \omega_q$. – оцениваемые коэффициенты.

Модели скользящего среднего MA дают прогноз значений функции Y_t на основе линейной комбинации ограниченного числа q остатков, в то время как авторегрессионные модели AR дают прогноз значения Y_t на основании линейной функции аппроксимации ограниченного числа p прошлых значений Y_t.

Использование понятия скользящего среднего в данном случае означает, что отклонение зависимой переменной от своего среднего, т.е. величина $Y_t - \mu$, является линейной комбинацией текущих и прошлых значений вектора случайных возмущений.

Модели с авторегрессией и скользящим средним

Авторегрессионную модель и модель со скользящим средним можно скомбинировать. При описании подобной комбинации используется обозначение $ARMA(p, q)$, где p – порядок авторегрессионной части модели, q – порядок части скользящего среднего. Модель $ARMA(p, q)$ имеет общий вид

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + +\epsilon_t - \omega_1 \epsilon_{t-1} - \omega_2 \epsilon_{t-2} - ... - \omega_q \epsilon_{t-q}$$

Данная модель позволяет составлять прогноз, зависящий как от текущего и прошлого значений зависимой переменной, так и от текущих и прошлых значений величины случайного возмущения.

Модели скользящего среднего MA дают прогноз значений функции Y_t на основе линейной комбинации ограниченного числа q остатков, в то время как авторегрессионные модели AR дают прогноз значения Y_t на основании линейной функции аппроксимации ограниченного числа p прошлых значений Y_t.

Использование понятия скользящего среднего в данном случае означает, что отклонение зависимой переменной от своего среднего, т.е. величина $Y_t - \mu$, является линейной комбинацией текущих и прошлых значений вектора случайных возмущений.

2.2 Тест Дики-Фуллера

Тест Дики — Фуллера (DF-тест, Dickey — Fuller test) — это методика, которая используется для анализа временных рядов для проверки на стационарность. Является одним из тестов на единичные корни.

Временной ряд имеет единичный корень, или порядок интеграции один, если его первые разности образуют стационарный ряд. Это усло-

вие записывается как $y_t \sim I(1)$ если ряд первых разностей $\Delta y_t = y_t y_{t_1}$ является стационарным $\Delta y_t \sim I(0)$.

При помощи этого теста проверяют значение коэффициента a в авторегрессионном уравнении первого порядка AR(1) $y_t = a\Delta y_{t_1+\epsilon t}$, где y_t — временной ряд, а ϵ — ошибка.

Если a = 1, то процесс имеет единичный корень, в этом случае ряд y_t не стационарен, является интегрированным временным рядом первого порядка — I(1).

Если |a| < 1, то ряд стационарный — I(0).

Для финансово-экономических процессов значение |a| > 1 не свойственно, так как в этом случае процесс является «взрывным». Возникновение таких процессов маловероятно.

Приведенное авторегрессионное уравнение AR(1) можно переписать в виде:

$\Delta y_t = b\Delta y_{t_1+\epsilon_t}$, где $b = a1$, а Δ — оператор разности первого порядка $\Delta y_t = y_t - y_{t_1}$.

- Нулевая гипотеза: H_0: b = 0 -— процесс нестационарен
- Альтернативаная гипотеза: H_1: b < 0 — процесс стационарен

2.3 Модели типа ARIMA и методология Бокса-Дженкинса

Бокс и Дженкинс предложили выделить класс нестационарных рядов, которые взятием последовательных разностей можно привести к стационарному виду типа ARMA. Если ряд после взятия d последовательных разностей сводится к стационарному, то для прогнозирования его уровней можно применить комбинированную модель авторегрессии и скользящего среднего, обозначаемую как ARIMA(p,d,q). Сокращение I в данной аббревиатуре означает «интегрированный». Методология Бокса-Дженкинса подбора ARIMA-модели для конкретного ряда наблюдений состоит из четырех этапов:

– идентификация модели – процесс выбора модели, в наилучшей степени соответствующей рассматриваемому реальному процессу;

– оценивание модели – использование регрессионных методов для получения оценок параметров, включенных в модель;

- тестирование модели.

– использование модели для прогнозирования.

2.4 AIC - информационный критерий Акаике

Информационный критерий — применяемая в эконометрике мера относительного качества эконометрических моделей, учитывающая степень "подгонки"модели под данные с корректировкой на используемое количество оцениваемых параметров. Т.е. критерии основаны на некотором компромиссе между точностью и сложностью модели.

Критерии различаются тем, как они обеспечивают этот баланс. информационные модели используются исключительно для сравнения моделей между собой, без содержательной интерпретации значений этих критериев. Обычно чем меньше значения критериев, тем выше относительное качество модели.

AIC (an information criterion) — информационный критерий Акаике - критерий, применяющийся исключительно для выбора из нескольких статистических моделей. Связан с концепцией информационной энтропии и расстоянии Кульбака-Лейблера, на основе которой был разработан этот критерий.

В общем случае AIC:

$$AIC = 2k - 2ln(L)$$

, где k — число параметров в статистической модели, и L — макси мизированное значение функции правдоподобия модели.

Предположим, что ошибки модели нормально и независимо распределены. Пусть n — число наблюдений и RSS-

$$RSS - \sum_{i=1}^{n} \epsilon_i^2$$

остаточная сумма квадратов. Далее мы предполагаем, что дисперсия ошибок модели неизвестна, но одинакова для всех них. Следовательно:

$$AIC = 2k + n[ln(2\pi RSS/n) + 1]$$

В случае сравнения моделей на выборках одинаковой длины, выражение можно упростить, выкидывая члены зависящие только от n:

$$AIC = 2k + n[ln(RSS)]$$

3 Используемые библиотеки

1. **Pandas**
   - Для работы с временными рядами
   2. **Matplotlib.pylab**
   - Для построения графиков функций
   3. **Statsmodels**
   - Используются статистические функции
   - Для постороения модели ARIMA
   4. **Sklearn.metrics**
   - Для R2 score

4 Реализация программы

1) Считываются данные из файла training.xlsx при помощи функции read_excel() из библиотеки pandas. По полученному временному ряду строим графики его скользящего среднего и стандартного отклонения. На основе этих графиков, а также теста Дики-Фулера (функция adfuller() из библиотеки statsmodels) оцениваем, является ли ряд стационарным.

2) Раскладываем временной ряд на тренд, сезональность и остаток в соответствии с аддитивной и мультипликативной моделями, при помощи функции seasonal_decompose из библиотеки statsmodels. Строим вывод на основании полученных результатов.

3) Ищем индекс интегрируемости ряда (функции diff() и dropna()), а также с помощью теста Дики-Фулера проверяем, является ли ряд стационарн ым.

4) С помощью функций автокорреляции и частичной автокорреляции (acf() и pacf()) подбираем нужные параметры для построения модели ARIMA.

5) Далее строим ARIMA-модель с помощью функии ARIMA(), после чего, с помощью функции model.fit() прогоняем ARIMA-модель через фильтр Калмана и с помощью model.predict() строим указания на тестовой выборке. Все эти функции принадлежат библиотеке statsmodels.

6) Функция r2_score() подсчитывает r2 score между реальной выборкой и предсказанной.

7) Функция model.aic() выводит значения критерия Акаике.